Rituerto-González, E., Mínguez-Sánchez, A., Gallardo-Antolín, A. y Peláez-Moreno, C. (2019). Data Augmentation for Speaker Identification under Stress Conditions to Combat Gender-Based Violence. *Applied Sciences*, 9(11), 2298.

DOI: https://doi.org/10.3390/app9112298

applied sciences

MDPI

Article

Data Augmentation for Speaker Identification under Stress Conditions to Combat Gender-Based Violence †

Esther Rituerto-González *, Alba Mínguez-Sánchez, Ascensión Gallardo-Antolín and Carmen Peláez-Moreno

Group of Multimedia Processing, Signal Theory and Communications Department, University Carlos III Madrid, 28911 Leganés, Madrid, Spain; minguezalba@gmail.com (A.M.-S.); gallardo@tsc.uc3m.es (A.G.-A.); carmen@tsc.uc3m.es (C.P.-M.)

* Correspondence: erituert@ing.uc3m.es

† This paper is an extended version of our paper published in the IBER Speech Conference, Barcelona, Spain, 21–23 November 2018.

Received: 1 March 2019; Accepted: 24 May 2019; Published: 4 June 2019

Abstract: A Speaker Identification system for a personalized wearable device to combat gender-based violence is presented in this paper. Speaker recognition systems exhibit a decrease in performance when the user is under emotional or stress conditions, thus the objective of this paper is to measure the effects of stress in speech to ultimately try to mitigate their consequences on a speaker identification task, by using data augmentation techniques specifically tailored for this purpose given the lack of data resources for this condition. An extensive experimentation has been carried out for assessing the effectiveness of the proposed techniques. First, we conclude that the best performance is always obtained when naturally stressed samples are included in the training set, and second, when these are not available, their substitution and augmentation with synthetically generated stress-like samples improves the performance of the system.

Keywords: speaker identification; emotions; stress conditions; data augmentation; synthetic stress

1. Introduction

In this paper, we analyze how stress affects speaker identification rates to determine if there is a significant difference when comparing it to a speaker identification system operating in neutral conditions. We aim at finding techniques to improve speaker identification systems when facing stressed speech, either by neutralizing the effects of stress or by training the system to cope with it. We propose data augmentation techniques both statistical and using synthetically generated speech under stressed conditions together with an analysis of the best feature extraction methods to design a stress-robust system [1].

The rest of the paper is organized as follows: Section 1 provides the context of this research explaining our view on how technology can be used to combat gender violence and highlights its motivation. Section 2 describes the state of the art on the subject of speaker identification and discusses features and classifiers used in the literature. Section 3 explains the architecture of the proposed system, Section 4 refers to the experimental set-up and results. Finally Section 5 outlines the conclusions and future work.

1.1. Gender-Based Violence

Violence against women is one of the biggest social problems in the world. Its cultural origin has made it an invisible phenomenon society is used to tolerate. In reality, one in three women globally

have faced sexual violence and at least one in five women are assaulted in college campuses in the United States alone [2].

Women, just because they are women, suffer from many different types of violence from their partners or their environment, that can go from controlling their decisions to physical and sexual aggressions. This violence affects, to a greater or lesser degree, women of any age, religion, economical and social conditions; it takes place at their own homes, at the workplace and even in public places.

To eradicate violence against women we need to take action against it, such as providing help and resources to women who suffer it, making prevention campaigns, training monitors and educating future generations in schools. Despite the universal need for women's safety, basic emergency reporting and response networks do not exist in much of the world, many nations do not even have a universal emergency access phone number that victims can call to report a crime.

A technological solution that allowed communities to rapidly respond to threats or assaults against any of its members, would ensure that help is always available. The project this research paper takes part in is born from the need to use technology to help finding a solution for this worldwide problem, using speech technologies and machine learning—among other disciplines—, with the ultimate goal of abolishing violence against women and achieving a fair and egalitarian society.

1.2. UC3M4Safety and Bindi

The research shown on this paper is part of a system called Bindi, a smart solution for women's safety by the UC3M4Safety group [3]. The UC3M4Safety is a multidisciplinary team for detecting, preventing and combating violence against women from a technological point of view at University Carlos III Madrid, Spain. The goal of this project is to develop a wearable solution for detecting through physiological sensory data, speech and audio analysis, and machine-learning algorithms, when a user is under an intense emotional state, such as panic, fear or stress, caused by a gender-based violence situation [4].

As shown in Figure 1, the Bindi system consists of two wearable devices and a smartphone application. One of the wearables, the bracelet, is in charge of measuring biometrics such as heart rate, galvanic skin response and temperature [5,6] through physiological sensors. The second device is a pendant, which incorporates a panic button and a microphone that records audio and speech, presumably from the user who is wearing it but also from the environment and other speakers. The bracelet is constantly registering and analyzing physiological data to identify any significant deviations with respect to the basal state. When an abnormality is detected, a trigger is sent to the pendant to start a microphone recording. This information is collected and sent via Bluetooth to the smartphone equipped with a software application that after examining and interpreting the audio signal—detecting sound events, voice levels, distinctive noises, etc.—under certain conditions, sends an alert to a group of people, previously selected by the victim, or to emergency services in order to alert of a threatening situation. The audio recorded is then stored in an only-read Cloud with several encrypting processes and accessibility restrictions.

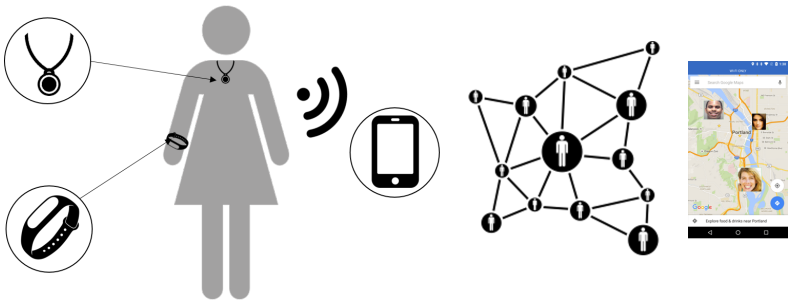

Figure 1. Conceptualization of the Bindi System.

Regarding speech, there are two main tasks to be performed within Bindi, Stress Detection [7] and Speaker Identification [1]. The former focuses in the detection of stress in the victim's voice whereas the latter relies on the proper identification of the victim despite the emotional conditions that may be present in her voice. These two tasks are interdependent as one needs from the help of the other. In this paper, we direct our attention to the latter, the development of a stress-robust speaker identification system.

Developing a personalized speaker identification module would allow the system to identify whether the voice belongs to the user or to someone else. Moreover, a system which is robust to stress could be able to recognize more accurately the main speaker when her voice exhibits stress conditions. Thus, this research is focused on speaker identification to recognize who does the voice captured by the microphone belong to even under emotional or stress conditions. This is further used to allow speaker adapted voice commands only obeying the victim but not the aggressor or even as judicial evidence.

1.3. Technological Challenges

In recent years, the interest in detecting and interpreting emotions in speech as well as synthesizing emotional speech has grown in parallel for a variety of applications. The research work done on emotions in speech is very extensive [8]. Speech Emotion Recognition (SER) consists of the identification of the emotional content of speech signals, the task of recognizing human emotions and affective states from speech. In the SER field, there are three important aspects being studied and discussed in the literature: the choice of suitable acoustic features [9], the design of an appropriate classifier [10] and the generation of an emotional speech database [11–13]. Some works propose multimodal approaches combining visual and speech data to improve and strengthen emotion recognition systems [14,15]. It is also well attested that speech recognition systems function less efficiently when the speaker is in an emotional state [16]. On the other hand, some works synthesize emotions in speech by systematically manipulating some of the parameters of human speech [17]. We can say then, that there is an important body of work about the effects of emotions in Automatic Speech Recognition (ASR), classification of emotions from speech or emotional speech synthesis, but the literature is scarce on the effects of emotions in Speaker Identification (SI) [18] and even less abundant about the influence of stress specifically.

One of the problems in this type of systems is the availability of databases recorded with emotional and neutral speech. Those existing are either recorded by actors simulating speech under those emotions, or by people to whom different emotions have been induced. This last option is truly complicated to implement—especially for negative emotions—and as a consequence, there are very few databases in which stressed speech is either simulated or recorded under real conditions, such as SUSAS [19], VOCE [20] or UT-Scope [21]. The main difficulty with these data relies on the labelling process.

Moreover, stress is not considered a proper emotion, although it is intimately related to anxiety and nervousness. It can be defined as a state of mental or emotional tension resulting from adverse or demanding circumstances. Stress may be induced by external factors (workload, noise, vibration, sleep loss, etc.) and by internal factors (emotion, fatigue, etc.). Among the physiological consequences of stress are respiratory changes, increased heart rate, skin perspiration and increased muscle tension of the vocal cords and vocal tract. All of these factors may, directly or indirectly, adversely affect the quality of speech [22] and help us discriminate between stressed or neutral speech using machine learning algorithms.

1.4. Contributions

The main contribution of this paper is the design of a Speaker Identification system robust to stress conditions in speech, by means of data augmentation techniques, that takes into account Bindi's computational restrictions and audio input characteristics. As mentioned before, the problem of detecting stress in speech is out of the scope of this paper.

2. Speaker Identification Related Work

Speaker Recognition is the automatic detection of persons from the characteristics of their voices, that is, by using voice biometrics [23]. Within SR, we can distinguish two tasks, Speaker Identification (SI) and Speaker Verification (SV). The former refers to the recognition of a particular user among a known number of users (a multiclass setting) and the latter aims to identify one specific user versus the rest of speakers (binary setting). In this paper we are assessing the influence of stress on a speaker identification system to further improve its performance with data augmentation techniques.

2.1. Feature Extraction

SI systems try to use acoustic features that differ between individuals to discriminate among them. Some of the features that exhibit good performance when used in neutral or emotionless conditions in speech-based systems are the Mel-Frequency Cepstral Coefficients (MFCC) [24]. Other prosodic features are widely used as well, such as intonation, stressed syllables and rhythm; and phonetic features, such as detection of the phones and their statistics. Finally, Linear Prediction is used in audio and speech processing for representing the spectral envelope of speech signals in compressed form. It is also a powerful speech analysis technique to provide estimates of speech parameters like pitch, duration and energy [25].

In speech emotion recognition, the current state of the art focuses on the use of formant frequencies, their bandwidths, pitch, log-energy and the so-called Normalized First-Order Autocorrelation Coefficients [26], among other features. Although for speaker identification under stress conditions there is hardly any previous work, MFCCs [27] together with prosodic features as the pitch, energy and word duration [22] have been used with successful results [28].

Beyond the previously mentioned hand-crafted features, automatic features extracted from raw data by DNN (Deep Neural Networks) is a successful trend achieving very innovative results [29,30]. Nevertheless, in our case, the use of complex DNN approaches is not currently possible due to their high computational load, delay and the availability of sufficiently big training datasets, which are three very important limitations within the Bindi system.

2.2. Data Augmentation

Data Augmentation (DA) is a key ingredient of the state of the art systems for image and speech recognition as it is a common strategy adopted to increase the amount of training data. It can also act as a regularizer to prevent overfitting [31] and to improve the performance in imbalanced class problems [32], making the whole process more robust. Due to the scarcity of data we mentioned in Section 1.3, this is a very good match for our case as the database we are using is quite small and has a noticeable imbalance. By using DA, we can increase the amount of data available and deal with the lack of balance between classes [33].

As for the SI system to be designed, it should be adapted to what we expect Bindi to find in a real world situation: the goal of our system is to detect the users speaking even when their voices present stressed conditions. For this reason, we would be facing a mismatch learning problem in which we may only have neutral utterances available for training gathered in an initial Bindi setup—given that the possibility of forcing the user to be stressed is difficult—whereas the real environment operating conditions would contain both neutral and stressed samples together. Our approach to solve this problem relies in the use of DA techniques to artificially synthesize stressed samples from neutral speech by applying slight modifications, and then adding these new synthetically stressed samples of speech to the training set, to ultimately build a stress-robust SI system.

2.3. Classifiers

Algorithms such as Gaussian Mixture Models (GMM) are generally employed for speaker recognition [34] and Support Vector Machines (SVM) are also widely applied [35,36]. Other studies

suggest the use of Deep Neural Networks (DNN) for speaker recognition [37,38]. However in this research, we aim to keep a balance between computational complexity and accuracy due to the hardware constraints of the device, where the battery consumption is critical and the scarse amount of training data originally available. After some preliminary tests to compare GMM, SVM and Multi-Layer Perceptron (MLP) classifiers for our task, the later—a precursor of Deep Neutral Networks with only one hidden layer—was chosen due to its simplicity, speed and better performance.

3. Methods

The block diagram for the training phase of our system is represented in Figure 2. The characteristics of the database employed, the automatic labelling process based on heart rate measurement, the two stages feature extraction, the data augmentation techniques and the normalization applied are described in detail in the next subsections. As for the test, the same methods are employed with the exception of the Data Augmentation block.

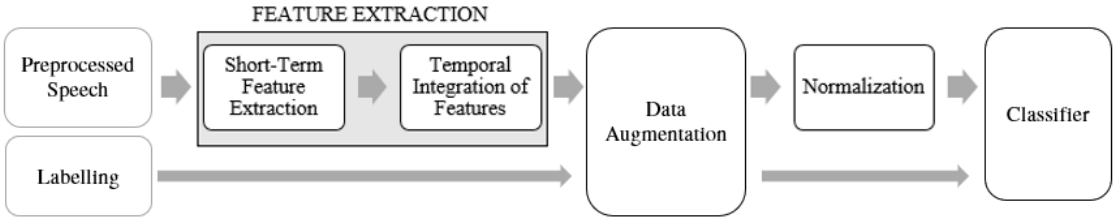

Figure 2. Block diagram of the system.

3.1. Corpus Database

Due to the difficulties for obtaining labels for the stress vs neutral conditions with Bindi we did not record audios using Bindi's system and microphone for this research. We chose the VOCE Corpus Database [20] instead which was not originally designed for speaker identification. There are several reasons why this database was chosen: first, for having data taken in real stress conditions; second, for offering data from sensors similar to those present in Bindi's bracelet from which heart rate measurements could be obtained; and third, due to the existence of previous studies [7] confirming the feasibility of relating heart rate metrics with stress in speech.

The last updated version of this dataset includes a total of 135 voice recordings that result from a set of 45 students (21 men, 17 women and 7 unidentified) from the University of Porto, with ages between 19 and 49 years. These voice files correspond to three different recording settings: *pre-baseline*, reading a standard text at least 24 h before the event in public (Public Speaking, PS); *baseline*, reading the same text as in the *pre-baseline* setting approximately 30 min before the PS; and *recording*, free speaking from a public event of free duration.

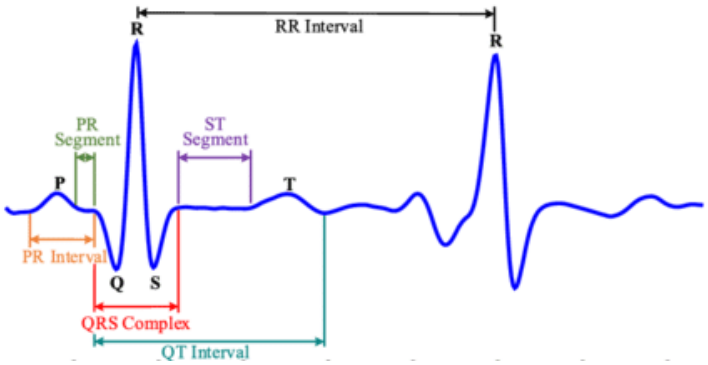

Figure 3. R peaks in ECG signal [39].

Together with these audio files, 117 files containing 2 measured physiological variables are provided and used to estimate the Heart Rate (HR). These measurements, taken with a Zephyr HxM BT2 device, are (i) Z_{ecg} representing an averaged and filtered HR value with a sampling period of 1 s. and (ii) Z_{ts} values that refer to the instants of time in which R peaks (shown in Figure 3) occur in the electrocardiogram obtained with the device, measured with an internal clock of 16 bits. Each of these values is accompanied by the Universal Time Coordinated (UTC) corresponding moment.

Furthermore, the database contains a metadata file that includes gender, age, health information, experience in public speaking, STAI (State-Trait Anxiety Inventory) [40] test scores and information about the quality of the recordings (energy level, saturation ...). Unfortunately, this is only provided for 38 out of the 45 individuals in the database and the database only gathers complete information (the 3 audio files and its corresponding HR values) from 21 individuals.

We decided to divide these 21 speakers into two sets, *Set 1* was composed of 10 speakers whose HR was coherent with the recordings in the sense that, when a speaker was reading the heart rate remained stable, but on the public speaking setting the HR rose significantly. *Set 2* was made out of the other 11 remaining speakers.

As for our specific application, the emotional conditions perceived in speech during an assault situation to be detected by Bindi—such as panic, anxiety or fear—, are not the same to stress conditions caused by public speech, but it seems to a recognizable prior condition that can lead to more intense emotions. As further work, we aim at using data specifically captured in these kinds of situations in the future.

3.2. Data Preprocessing

The audio recordings from VOCE were converted from stereo to mono signals to ease their handling. Also, we performed a downsampling from 44,100 Hz to 16,000 Hz to reduce the battery consumption since the transmission of audio from the bracelet to the smartphone is very costly in these terms. We continue by normalizing the signals to have zero mean and fall withing the $[-1, 1]$ range. As a final step, the signals went through a Voice Activity Detector module (VAD) [41]. This VAD algorithm is designed for improving speech detection robustness in noisy environments, by removing one-second length chunks of non-speech audio from which no decision about stress or speaker can be taken. Each of these transformations are performed for speech audio processing in Bindi as well.

As for the Heart Rate measures collected in the database, the original signed Z_{ecg} values were converted to unsigned ones from 0 to 255. The Zts sequences were discarded since they were considered too noisy and Z_{ecg} already provided the HR information needed with a reasonable temporal resolution.

3.3. Automatic Stress Labelling

In this paper we work with two types of labels for each audio utterance: boolean *stress labels* that indicate the presence of stress and *speaker labels*, taking values from 1 to n, representing the speaker id of each the audio sample, being n the total number of speakers.

Labelling a speech signal to determine stress presence is a delicate matter, since there is not a prescribed way to do so, stress is non binary and its determination subjective. Instead of the labels included in the original VOCE Corpus to each recording situation (0 for the full *pre-baseline* or *baseline* sequences and 1 for *recording*) we generated the labels from the HR sequences. Every one-second long audio frame is labelled as stressed or neutral using a speaker dependent HR threshold established for each of the speakers using their respective *pre-baseline* recordings. Two different HR thresholds were compared: the *pre-baseline* HR average plus the standard deviation and the 75% percentile of the HR values.

Following the work done by A. Mínguez-Sánchez [7] we used the latter, given the automatic stressed classifiers implemented gave better results. Table 1 provides the number of samples per recording condition. Each sample represents one-second length audio frames with approximately 3.5 h of non-silent speech.

Table 1. Number of samples.

Set Label	Neutral	Stressed	Total
Set 1	1389	3989	5378
Set 2	1716	4858	6574
Total	3105	8847	11,952

3.4. Feature Extraction

After the speech pre-processing block, the signal is ready for feature extraction. The acoustic features of speech extracted from audio signals should reflect both anatomy (e.g., size and shape of the throat and mouth) and learned behavioral patterns (e.g., voice pitch, speaking style). For each of them, a window of 20 ms with 50% overlapping is used. The reason these are the values chosen it is because they fall within the range of standard values used to analyze the temporal evolution of the signals [9]. To convert the feature vectors into one second resolution values alignable with the labels explained in the previous subsection, the mean and standard deviation of the acoustic characteristics over segments of one second length are computed, achieving one feature vector per second of audio. These features are summarized in Table 2 and were selected according to the literature for emotions and speaker recognition [9,42].

Table 2. Components of the feature vectors.

Row Index	Feature
0–12	Mean MFCC
13–25	Standard Deviation MFCC
26–28	Mean first three formants
29–31	Standard Deviation first three formants
32	Mean Pitch
33	Standard Deviation Pitch

3.5. Data Augmentation

Although in real speech we may encounter different emotional conditions, attitudes or behaviours in the speaker that may affect the produced speech, we only focus on the existence of neutral and stressed utterances in this study for simplicity purposes due to its preliminary character.

Due to the low number of samples of the database and the fact that the user registration would presumably lack stressed speech, we considered the generation of a synthetically stressed database of utterances using data augmentation. To be able to produce stressed speech out of neutral utterances we carried out an informal analysis by listening the audio signals provided in the database, which were initially classified as stressed and neutral speech signals and took note of the differences that could be appreciated between them. Secondly, we measured objectively those differences between stressed and neutral utterances of the same speakers in terms of locution speed and pitch.

As a first outcome, we realized that the locution speed may reflect the stress of a person, as people tend to pronounce more words per second and produce longer pauses when they are stressed. In stressed conditions there is also a tendency to rise the pitch. Therefore, the speaking rate and pitch of the speaker are two variables to be modified in order to artificially simulate speech under stress conditions. The Python SOX library was used for this purpose [43].

3.6. Normalization

A standard zero mean and variance normalization is performed to avoid attributes with greater numeric ranges to dominate those in smaller numeric ranges.

4. Experiments

In this section, we detail the experimentation carried out along with the results to evaluate the performance of the speaker identification under stress conditions.

4.1. Balanced Data

The fact that the data instances were not balanced, i.e., there are speakers with significantly more samples than others, led us to perform an adjustment for each set and condition to get consistent estimates, give all classes—in this case, speakers—need to be seen as equally important from the point of view of the classifier and minimize loss in the training phase. Nevertheless, the use of a purely statistical over-sampling technique would have a big drawback in our case since the imbalance is very severe and the amount of artificial data created would be too large. To cope with this problem, we first under-sampled the set of neutral data admitting a maximum of 120 samples per speaker in both sets (1 and 2) as well as the stressed set using a threshold of 300 samples. Applying an over-sampling technique (in particular, SMOTE for Python) [32] to the under-sampled data resulted in sufficient new samples achieving a balanced data set but without including a disproportionate amount of artificial data.

4.2. Match and Mismatch Conditions

Originally, for an initial experimental set-up we used the data available for Sets 1 and 2 together (21 speakers). This preliminary experiment is made to observe the behaviour of the speaker identification rate in mismatch conditions. First of all, we divided the data into Neutral (N) and Stressed (S) speech and experimented training with one type of speech, testing with the other and then mixing both types. The results in terms of accuracy—the percentage of audio segments correctly classified—can be found in Table 3. In order to get reliable results these experiments were repeated 50 times, where in each repetition the data used for testing was chosen randomly, excluding samples used for training.

Table 3. Results for matched and mismatched settings.

Training Set	Test Set	Mean (%)	Std (%)
Neutral	Neutral	96.73	0.33
	Stressed	79.21	0.90
Stressed	Stressed	95.87	0.28
	Neutral	90.89	0.49
Mixed	Mixed	96.05	0.12

As a first conclusion, matched settings are better than mismatched ones as expected. When training with neutral utterances and testing with stressed ones, accuracy decreases in more than a 15% with respect to match settings, so it seems proved that stressed speech does have different characteristics compared to neutral speech that affect SI. On the contrary, when training with stressed utterances of speech and testing with neutral ones, the decrease in accuracy with respect to the matched setting is not that important (5%) comparing it to the reversed case, leading us to think that stressed speech could be sparse data in which neutral speech could be contained but not vice versa. About the mixed conditions experiments, the accuracy reached a 96.05%, achieving a good result for this particular task.

4.3. Pitch and Speed Modifications

The pitch and the elocution speed were two variables we informally observed to be changing between neutral and stressed speech. As a consequence, we performed an analysis to measure the differences between the mean pitch from neutral to stressed audio frames for each speaker

using VoiceBox [41]. An estimation of the average elocution speed for each user was calculated as well, computing the mean number of words per second of each speaker by obtaining an automatic transcription of each of the recordings using Google Speech Recognition [44] and dividing it over the length of the audio signals after having removed silent audio frames with the VAD module.

The differences in pitch from neutral to stressed speech were in a range between a relative percentage of −2% and +7% for all speakers, increasing on average 2.2%. In regard to the elocution speed, subjectively, it seems to rise in stressed speech, however our analysis gave us the opposite conclusion. The number of words per second was higher when the user was reading a text, 2.2 words/s on average, in comparison with when the speaker was performing an oral presentation, 1.85 words/s. By listening to the signals, we determined that the words were pronounced faster but there were many short pauses and pause fillers—words like 'ehm', 'um', 'ah'—in between them, that did not count as words for the transcription but were not removed by the VAD either. Those causes lead to an overall lower elocution rate.

Thus, we applied modifications in the locution speed and the pitch on the original database, to produce synthetically stressed samples of speech. The pitch was modified by the following relative percentages [−6%, −3%, +3%, +6%] and the speech signals were slowed down—with the aim of extending the duration—by the following percentages [−20%, −15%, −10%, −5%]. All of these modifications were applied to the original audio signals and resulted—in what we would name—a new synthetically stressed set per modification. In this manner, we augmented our data by a factor of 9, the original dataset plus 8 modifications. Figure 4 presents a block schematic of the original data and the synthesized one. Synthetic Stress (SS) and Synthetic Super Stressed (SSS) represent the synthetically stressed collections obtained from N and S sets respectively.

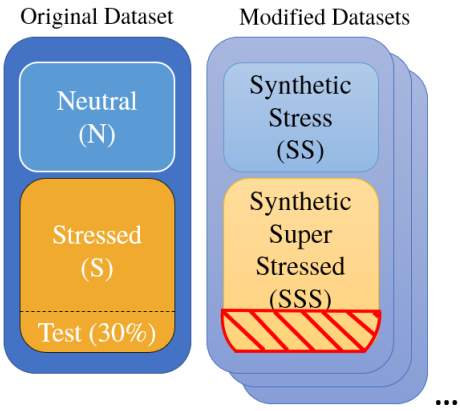

Figure 4. Schematic of Original and Modified Datasets. On the left side, we represent in a block basis the original dataset, composed by neutral and stressed samples. In this case, we used the 30% of the examples of the the stressed collection as the Test set for later experiments. On the right, we represent a diagram of one of the synthetically stressed sets, where the 30% of data used before as Test was removed to obtain more reliable results. There are several synthetic datsets, one per modification applied.

4.4. Preliminary Experiments

In the next experimental set-up, we aim at measuring the accuracy achieved by the system when training with the different pitch and speed modified sets and testing with originally stressed utterances for the first set of speakers, Set 1. The results achieved in these experiments should reflect which modification imitates best the original stressed samples. We kept the test set fixed for these experiments, a 30% of the samples of original stressed speech. Additionally, the same 30% in every synthetic super-stressed set was removed to achieve a more accurate comparison between experiments and guarantee that the test samples were never present in the training set even if they had been modified by our augmenting procedure. In Figures 5 and 6 we present the results obtained,

we enumerate the data used for the training step on the *X* axis, the *Y* axis represents the accuracy achieved and each colour bar indicates the modified set used for training.

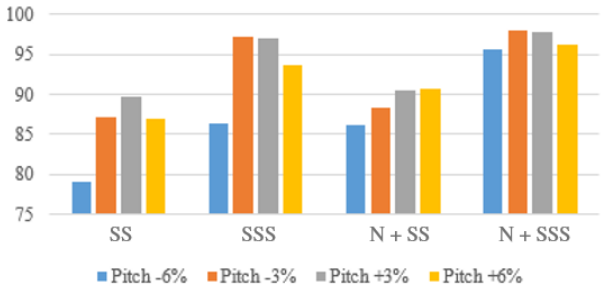

Figure 5. Results for synthetic datasets, Set 1 Pitch modifications.

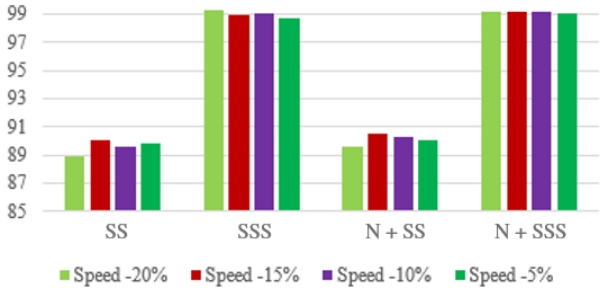

Figure 6. Results for synthetic datasets, Set 1 Speed modifications.

In Figure 5 we can observe that the modifications that obtain the highest accuracies are *Pitch +3%* and *Pitch −3%*. When it comes to Figure 6, although the speed results are very similar, the alteration that in general works worse is *Speed −20%*. As for the training sets used, the SSS set works better than the SS in both cases.

4.5. Synthetic Stress Combinations

For the next set of experiments we decided to perform the modifications to the audio recordings of Set 2 that had achieved higher accuracy rates in Set 1. These were pitch [−3%, +3%] and signal speed [−15%, −10%, −5%] as mentioned. We joined Sets 1 and 2, transforming the problem in a 21-speaker SI task and combined all the synthetic stress data into one dataset, augmenting in a factor of 6 the original data size, 5 modifications plus the original dataset. The same analysis were done for Set 1 and Sets 1 + 2.

The equivalence between each training set used and case number is shown in Figure 7. On this figure, we can appreciate the different combinations of configurations of the original and modified datasets used for training. These compositions were grouped forming different combinations in order to acknowledge the differences in accuracy for each particular setting used in the training stage. These experiments were repeated 20 times for reliability with the outcome shown in Table 4.

Case	1	2	3	4	5	6	7	8	9	10
Training Data	N	S	N + S	N + SS	N + SSS	SS	SSS	N + S + SSS	N + S + SS + SSS	N + SS + SSS

Figure 7. Equivalence between training data used and case of experiment in Section 4.5.

In Table 4 we can appreciate two types of experiments, some in which we substitute data and others where we augment data on the training stage. As for substituting the original set by a synthetically stressed one, we have experiments 6 and 7 to be compared with experiments 1 and 2

respectively. Data substitution achieves similar results to those with original data when using synthetic data converted out of neutral speech for training (case 1 vs. case 6) as well as better identification rates when using synthetic data obtained from stressed speech (case 2 vs. case 7).

Table 4. Extensive experimentation results.

Case	Set 1 Mean	Set 1 Std	Set 1 + 2 Mean	Set 1 + 2 Std
1	89.71	0.56	78.55	0.60
2	98.59	0.16	97.37	0.21
3	98.48	0.23	97.21	0.26
4	89.97	0.39	80.46	0.53
5	99.93	0.05	99.16	0.11
6	89.72	0.53	78.19	0.71
7	99.88	0.07	99.21	0.13
8	99.91	0.07	99.45	0.08
9	99.94	0.06	99.22	0.11
10	99.91	0.07	98.97	0.14

Data augmentation experiments are 3, 4, 5, 8, 9 and 10. The outcome is indeed positive, the best results are achieved in experiment 8 with a 99.45% of accuracy for Sets 1 + 2. These results show us that augmenting the data with synthetically stressed utterances of speech boosts the SI rate.

One of the objectives of these experiments was to determine wether experiment 4 could outperform experiment 2. This would mean that we had accomplished the task of generating appropriate synthetically stressed speech out of neutral utterances. However, we can see that the procedure we employed was not enough to be used as a substitute. Nevertheless, in Table 4 one can see that case 4 performs better than case 6, which in turn outperforms case 1. This shows that stressing speech synthetically and using it as training data alongside with original stressed data increases the performance of the SI system.

5. Conclusions and Future Work

In this research, our goal was to analyze how stressed speech influences the Speaker Identification systems performance. We have identified a problem, stressed speech in the testing stage affects negatively when SI systems are trained only with neutral speech.

As for the case of match and mismatch conditions, in the mixed setting—using neutral and stressed original utterances—the SI system achieves a 96.05% of accuracy, a satisfactory rate for this type of tasks, demonstrating that the set of features chosen for the task is adequate.

In the preliminary experiments for data substitution, depending on the difference between the synthetic data and the original one used for training, some substitutions outperform the results achieved by original data. Besides, the modifications over the pitch of the speaker work better when we include synthetically stressed samples for training, than when we include the modifications in speech speed. However, when we use super synthetic stressed samples for training, the sets modified by changes in speed achieve better results.

Regarding the experiments for augmenting the database with artificial stress, we can conclude that the generation of different synthetically stressed utterances of speech by modifications in pitch and speed, and their addition to the database, enlarges meaningfully the instances to work with, improving substantially the results achieved by the Speaker Identification system with a 99.45% of accuracy.

5.1. Future Work

Several experiments and methods remained unexplored and were left for future work:

- Our target in this research is a Speaker Identification task, a multiclass problem. However, the objective of the device to be built in Bindi is a Speaker Verification system, thus the next

step would be to transform the system into a binary setting. These two approaches are not straightforwardly comparable but we believe that the problems and solutions can be translated to one another.

- Although stress seems to be an emotional condition that usually precedes more intense emotions, we aim to find or record a database that includes emotions in speech during an assault situation—such as panic, anxiety or fear—to work with.
- Finding techniques to strengthen the system by degrading audios as if they had been recorded in a real environment. Perhaps to simulate real world situations in which the recorded voice is not clean, we could add noise to the same database used and analyze its effect, either by sounds recorded at outdoors and indoors atmospheres or with white/pink noise.
- Further analyzing the differences between neutral and stressed speech to find new modifications to be applied to neutral speech to transform it into appropriate synthetically stressed speech.
- Implementing new methods for recording stressed speech in the training phase using Bindi, such as Stroop Effect games [45] in which the speaker should experiment stress, would be a way to count with originally stressed samples in the training stage.
- Heart Rate Variability (HRV) is known to have a strong relationship with stress. To deepen into this question and further investigating, we could try to find correlations between HRV and stress, adding this other biometric feature to the stress-robust SI system.
- With the use of data augmentation techniques we have collected a much larger database and we could therefore employ more powerful Deep Learning algorithms in the future, provided the device employed is able to cope with it in real time.

Author Contributions: Methodology, software and experimentation, E.R.-G.; data curation, A.M.-S.; writing—original draft preparation; E.R.-G. and A.M.-S.. Conceptualization, methodology, supervision and writing—review and editing, A.G.-A. and C.P.-M.

Funding: This work is partially supported by the Spanish Government-MinECo project TEC2017-84395-P and Madrid Regional Project Y2018/TCS-5046.

Acknowledgments: The authors thank all of the members of UC3M4Safety for their contribution and support in the definition of requirements and constraints of the present work to be included as a part of the Bindi devices. The authors also thank the committee of the IV Edition of the Pilar Azcárate Gender Violence Research Awards at University Carlos III Madrid, Spain for the prize awarded to the inception of this research that has provided high motivation.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results'.

References

1. Rituerto-González, E.; Gallardo-Antolín, A.; Peláez-Moreno, C. Speaker Recognition under Stress Conditions. In Proceedings of the IBER Speech, Barcelona, Spain, 21–23 November 2018; pp. 15–19. [CrossRef]
2. World Health Organization. *Global and Regional Estimates of Violence against Women: Prevalence and Health Effects of Intimate Partner Violence and Non-Partner Sexual Violence*; World Health Organization: Geneva, Switzerland, 2013.
3. *UC3M4Safety—Multidisciplinary Team for Detecting, Preventing and Combating Violence against Women*; University Carlos III of Madrid: Leganes, Spain, 2017.
4. Miranda, J.A.; Canabal, M.F.; Portela-García, M.; Lopez-Ongil, C. *Embedded Emotion Recognition: Autonomous Multimodal Affective Internet of Things*; CEUR Workshop: Tenerife, Spain, 2018; Volume 2208, pp. 22–29.
5. Miranda, J.A.; Canabal, M.F.; Lanza, J.M.; Portela-García, M.; López-Ongil, C.; Alcaide, T.R. Meaningful Data Treatment from Multiple Physiological Sensors in a Cyber-Physical System. In Proceedings of the DCIS 2017: XXXII Conference on Design of Circuits and Integrated Systems, Barcelona Spain, 22–24 November 2017.

6. Miranda-Calero, J.A.; Marino, R.; Lanza-Gutierrez, J.M.; Riesgo, T.; Garcia-Valderas, M.; Lopez-Ongil, C. Embedded Emotion Recognition within Cyber-Physical Systems using Physiological Signals. In Proceedings of the DCIS 2018: XXXIII Conference on Design of Circuits and Integrated Systems, Lyon, France, 14–16 November 2018.
7. Minguez-Sanchez, A. Detección de Estrés en Señales de voz. Bachelor's Thesis, University Carlos III Madrid, Madrid, Spain, 2017.
8. Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* **2018**, *61*, 90–99. [CrossRef]
9. Anagnostopoulos, C.N.; Iliou, T.; Giannoukos, I. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artif. Intell. Rev.* **2015**, *43*, 155–177. [CrossRef]
10. Noroozi, F.; Kaminska, D.; Sapinski, T.; Anbarjafari, G. Supervised Vocal-Based Emotion Recognition Using Multiclass Support Vector Machine, Random Forests, and Adaboost. *J. Audio Eng. Soc.* **2017**, *65*, 562–572. [CrossRef]
11. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Provost, E.M.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
12. Haq, P.J.S. *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*; University of Surrey: Guildford, UK, 2014.
13. Vryzas, N.; Kotsakis, R.; Liatsou, A.; Dimoulas, C.; Kalliris, G. Speech Emotion Recognition for Performance Interaction. *J. Audio Eng. Soc.* **2018**, *66*, 457–467. [CrossRef]
14. Vryzas, N.; Vrysis, L.; Kotsakis, R.; Dimoulas, C. Speech Emotion Recognition Adapted to Multimodal Semantic Repositories. In Proceedings of the 2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Zaragoza, Spain, 6–7 September 2018; pp. 31–35. [CrossRef]
15. Kahou, S.E.; Bouthillier, X.; Lamblin, P.; Gülçehre, Ç.; Michalski, V.; Konda, K.R.; Jean, S.; Froumenty, P.; Dauphin, Y.; Boulanger-Lewandowski, N.; et al. EmoNets: Multimodal deep learning approaches for emotion recognition in video. *J. Multimodal User Interfaces* **2016**, *10*, 99–111. [CrossRef]
16. Hansen, J.H.; Patil, S. *Speaker Classification. Speech Under Stress: Analysis, Modeling and Recognition*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 108–137.
17. Murray, I.; Arnott, J.L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J. Acoust. Soc. Am.* **1993**, *93*, 1097–1108. [CrossRef]
18. Wu, W.; Zheng, F.; Xu, M.; Bao, H. Study on speaker verification on emotional speech. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.
19. Hansen, J.H.L. *Speech under Simulated and Actual Stress (SUSAS) Database*; Linguistic Data Consortium: Philadelphia, PA, USA, 1999.
20. Aguiar, A.; Kaiseler, M.; Cunha, M.; Meinedo, H.; Almeida, P.R.; Silva, J. VOCE Corpus: Ecologically Collected Speech Annotated with Physiological and Psychological Stress Assessments. In Proceedings of the LREC 2014: 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 26–31 May 2014.
21. Ikeno, A.; Varadarajan, V.; Patil, S.; Hansen, J.H.L. UT-Scope: Speech under Lombard Effect and Cognitive Stress. In Proceedings of the 2007 IEEE Aerospace Conference, Big Sky, MT, USA, 3–10 March 2007; pp. 1–7. [CrossRef]
22. Steeneken, H.J.M.; Hansen, J.H.L. Speech under stress conditions: Overview of the effect on speech production and on system performance. In Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ, USA, 15–19 March 1999; Volume 4, pp. 2079–2082. [CrossRef]
23. Poddar, A.; Sahidullah, M.; Saha, G. Speaker verification with short utterances: A review of challenges, trends and opportunities. *IET Biom.* **2018**, *7*, 91–101. [CrossRef]
24. Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B. Speaker Verification Using Adapted Gaussian Mixture Models. *Digit. Signal Process.* **2000**, *10*, 19–41. [CrossRef]
25. Campbell, J.P. Speaker recognition: A tutorial. *Proc. IEEE* **1997**, *85*, 1437–1462. [CrossRef]
26. Li, Y.; Zhao, Y. Recognizing Emotions in Speech Using Short-term and Long-term Features. In Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, 30 November–4 December 1998; Volume 6, pp. 2255–2258.

27. Senthil Raja, G.; Dandapat, S. Speaker recognition under stressed condition. *Int. J. Speech Technol.* **2010**, *13*, 141–161. [CrossRef]
28. Zheng, N.; Lee, T.; Ching, P.C. Integration of Complementary Acoustic Features for Speaker Recognition. *IEEE Signal Process. Lett.* **2007**, *14*, 181–184. [CrossRef]
29. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204. [CrossRef]
30. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Trans. Multimed.* **2018**, *20*, 1576–1590. [CrossRef]
31. Simard, P.Y.; Steinkraus, D.; Platt, J.C. Best practices for convolutional neural networks applied to visual document analysis. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, Edinburgh, UK, 3–6 August 2003; pp. 958–963. [CrossRef]
32. Bowyer, K.W.; Chawla, N.V.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2011**, *16*, 321–357.
33. Rebai, I.; BenAyed, Y.; Mahdi, W.; Lorré, J.P. Improving speech recognition using data augmentation and acoustic model fusion. *Procedia Comput. Sci.* **2017**, *112*, 316–322. [CrossRef]
34. Meuwly, D.; Drygajlo, A. Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM). In Proceedings of the A Speaker Odyssey—The Speaker Recognition Workshop, Crete, Greece, 18–22 June 2001; pp. 145–150.
35. Campbell, W.M.; Sturim, D.E.; Reynolds, D.A. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Process. Lett.* **2006**, *13*, 308–311. [CrossRef]
36. Abdalmalak, K.A.; Gallardo-Antolín, A. Enhancement of a text-independent speaker verification system by using feature combination and parallel structure classifiers. *Neural Comput. Appl.* **2018**, *29*, 637–651. [CrossRef]
37. Lei, Y.; Scheffer, N.; Ferrer, L.; McLaren, M. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 1695–1699. [CrossRef]
38. Yu, D.; Seltzer, M.L.; Li, J.; Huang, J.; Seide, F. Feature Learning in Deep Neural Networks—A Study on Speech Recognition Tasks. *arXiv* **2013**, arXiv:1301.3605.
39. He, R.; Wang, K.; Li, Q.; Yuan, Y.; Zhao, N.; Liu, Y. ; Zhang, H. A Novel Method for the Detection of R-peaks in ECG Based on K-Nearest Neighbors and Particle Swarm Optimization. *EURASIP J. Adv. Signal Process.* **2017**, *2017*, 82. [CrossRef]
40. Spielberger, C.D.; Gorssuch, R.L.; Lushene, P.R.; Vagg, P.R.; Jacobs, G.A. *State-Trait Anxiety Inventory (STAI)*; Consulting Psychologists Press: Palo Alto, CA, USA, 1968.
41. Brookes, M. Voicebox: Speech Processing Toolbox for Matlab [Software]. Available online: www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox (accessed on 4 June 2019).
42. Tirumala, S.S.; Shahamiri, S.R.; Garhwal, A.S.; Wang, R. Speaker identification features extraction methods: A systematic review. *Expert Syst. Appl.* **2017**, *90*, 250–271. [CrossRef]
43. Bittner, R.; Humphrey, E.; Bello, J. PySOX: Leveraging the Audio Signal Processing Power of SOX in Python. In Proceedings of the International Conference on Music Information Retrieval (ISMIR-16) Conference Late Breaking and Demo Papers, New York, NY, USA, 8–11 August 2016.
44. Zhang, A. Speech Recognition Library for Python (Version 3.8) [Software]. 2017. Available online: https://github.com/Uberi/speech_recognition (accessed on 4 June 2019).
45. Stroop, J.R. Studies of Interference in Serial Verbal Reactions. *J. Exp. Psychol. Gen.* **1992**, *121*, 15–23. [CrossRef]